The Journal of Transport and Land Use http://jtlu.org
Vol. 10 No. 1 [2017] pp. 675–694

JTLU

# Analyzing spatiotemporal congestion pattern on urban roads based on taxi GPS data

**Kaisheng Zhang**
State Key Laboratory of Ocean Engineering
Shanghai Jiao Tong University
zhangkssjtu@hotmail.com

**Daniel (Jian) Sun**
China Institute of Urban Governance
Shanghai Jiao Tong University
danielsun@sjtu.edu.cn

**Suwan Shen**
University of Hawaii, Manoa
suwans@hawaii.edu

**Yi Zhu**
Shanghai Jiao Tong University
yizhu@smart.mit.edu

**Abstract:** With the development of in-vehicle data collection devices, GPS trajectory has become a priority source to identify traffic congestion and understand the operational states of road network in recent years. This study aims to investigate the relationship between traffic congestion and built environment, including traffic-related factors and land use. Fuzzy C-means clustering was used to conduct an exhaustive study on the 24-hour congestion pattern of road segments in an urban area, so that the spatial autoregressive moving average model (SARMA) could be introduced to analyze the output from the clustering analysis to establish the relationship between built environment and the 24-hour congestion pattern. The clustering result classified the road segments into four congestion levels, while the regression explained the impact of 12 traffic-related factors and land-use factors on the road congestion pattern. The continuous congestion was found to mainly occur in the city center, and the factors, such as road type, bus station in the vicinity, ramp nearby, commercial land use, and so on, had large impacts on congestion formation. The Fuzzy C-means clustering is proposed to be combined with quantitative spatial regression, and the overall evaluation process will assist to assess the spatial-temporal levels of service regarding traffic from the congestion perspective.

**Keywords:** Congestion pattern, taxi GPS data, fuzzy C-means clustering, spatiotemporal regression, built environment factor

**Article history:**
Received: October 18, 2016
Received in revised form: April 13, 2017
Accepted: April 13, 2017
Available online: June 19, 2017

## 1 Introduction

In urban road network, the recurrent or current congestion of a certain road segment may largely impact the local network and reduce travel efficiency. Consequently, it is important to identify the


http://dx.doi.org/10.5198/jtlu.2017.980
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 3.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

congested road segments in real-time and implement corresponding traffic mitigation strategies. Fixed facilities, such as inductive loops, traffic surveillance systems and microwave radars are commonly used for road traffic detection and various data collection, including traffic speed, traffic volume, density and vehicle classification. However, such facilities are expensive and mostly only serve intersections or freeways. The sparse sensor network makes it difficult to identify the problematic links in real-time. Global Position System (GPS) data enables the simultaneous analysis of spatial and temporal patterns in traffic information. While Global Position System (GPS) data collected from vehicles and mobile phones has become increasingly popular, among which taxi GPS data is preferable because of its exhaustive coverage of the road network, high frequency and the acquisition convenience (Ding et al., 2014; Wang, Peng, Lu, Sun, & Bai, 2017). In general, no fixed definitions were proposed for the level of congestion, while travel speed is the most commonly used indicator for traffic congestion assessments (He, Yan, Liu, & Ma, 2016). This study intends to investigate the 24-hour congestion pattern of the road network by speed, to classify the road segments by their speed patterns through Fuzzy C-means (FCM) clustering, and to analyze the problematic segments with continuous low speed or unconventional congestion. Spatial models based on *geographical detector, MORAN's Index* and spatial regression (*SARMA*) were developed to analyze the relationship between congestion patterns and the surrounding built environment.

Compared with mobile phone data, floating car data, cargo transport vehicle record and navigation system, taxi GPS trace data is one of the easiest available sources for accurate travel route and travel time records for a wider area with more road details. Data mining based on taxi trip can be traced back to the 1970s (Goddard, 1970), which has been applied to a wide range of studies, mainly including activity-based and infrastructure-based fields. The activity-based studies mostly focus on driver behavior, supply-demand pattern, and traffic state analysis, while the infrastructure-based studies mainly focused on lanes channelization (Tang, Yang, Kan, & Li, 2015) and signal-timing estimation (Yu & Lu, 2016).

From driver behavior perspective, Zhang, Qiu, Duan, Du, and Lu (2015) proposed a space-time visualization method to demonstrate taxi daily trajectories by GIS-T to recognize working time, operating range, and residence location without time division. Qing, Parfenov, and Kim (2015) compared direct extracted datas like travel distance, speed, demand, and supply mismatch of taxi trip between fair weather and extreme storm using Manhattan GPS data, and discovered the reduction in trip distance and supply of drives during the extreme storm. Meanwhile, Hwang, Wu and Jian (2006) used structural equation modeling techniques to improve taxi dispatching service based on consumer preference modeling based on questionnaires and GPS data, however, the time variation was not considered. Tang, Jiang, Li, and Li (2016) analysis drivers' customer searching behaviors by proposed two-layer model based on GPS data, and path size, path distance, and travel time act as influencing factors, however, geographical factors like land use and traffic-related factors were not involved. Yazici, Kamga and Singhal (2016) studied New York taxi drivers' decisions on pick-ups or cruising for passengers after end of trips at the JFK airport using a logistic regression based on temporal and weather factors, among which peak hour was treated as an independent variable. Chen, Zhang, Li, and Zhou (2014) introduced B-planner for planning bidirectional night bus routes using taxi GPS traces, and conducted qualitative analyses using clustering results.

Taxi GPS data also contributes to supply-demand pattern of taxi. Hu, An and Wang (2014) analyzed time of day and day of the week variations in urban taxi drivers' service time and operation frequency by descriptive statistics. However, the authors failed to further construct the relationship between service and built environment. Qian and Ukkusuri (2015) combined geographically weighted regression (GWR) with NYC taxi data to establish the relationship between taxi ridership and demographic, land-use and transportation system. The hourly demand variation was not reflected as only the daily ridership was aggregated and analyzed. Lu and Li (2014) used taxi GPS data to predict OD

distribution, while the statistical methods still treated one-day data as a whole and explained by time series without multivariable combination.

Conventional studies based on GPS data regarding traffic state include congestion detection (Montero, Pacheco, Barcelo, Homoceanu, & Casanovas, 2016), link or route travel time, speed and distance measurements (Tulic, Bauer, & Scherrer, 2014; Jiménez-Meza, Arámburo-Lizárraga, & Fuente 2013); detecting urban road network accessibility (Cui et al., 2016). Most of these studies only focus on peak hours and only use descriptive statistics without deep analysis about influencing factors for time, speed, distance, etc. Twice transformation is deficient in state identification. Azimi & Zhang (2010) applied clustering algorithms (K-means, Fuzzy C-means and CLARA) to sort freeway traffic conditions by traffic flow, however, the result was qualitative, which was hard for application.

According to the literature review, the previous taxi GPS data based researches mainly have three problems as follows. First, a majority of studies only focus on peak hours or discrete hours, or even aggregated daily data without hourly division, which may only disclose sudden breakdown of peak hours, and fail to reflect hourly volatility of traffic system, or neglect temporal difference. Additional analyses generally focus on certain straightforward indices, such as speed, flow rate or travel time, while other indirect or consequent indicators, such as congestion, tend to acquire less attention because of devoid of fixed quantitative definition. Finally, the existing studies failed to utilize independent variables, such as land use, built environment, and traffic-related factors, to explain the information extracted from taxi GPS data, while GPS data is always used to identify those infrastructure-based factors, as well as land use as mentioned (Pan, Qi, Wu, Zhang, & Li, 2013).

This research aims to treat the speed pattern by 24-hour-based dataset, trying to reflect the volatility trend by clustering. An analytical framework combining clustering method and spatial regression is proposed to cover the shortage of twice transformation for congestion and quantitative analysis. Land-use variables and traffic-related factors are included in regression.

The research flow chart is presented in Figure 1: This paper first use taxi GPS data of Shanghai to classify road segments in an urban area based on their 24-hour congestion pattern using Fuzzy C-means clustering (FCM), which allocated objects to clusters by probability. We then set such probability classification as the dependent variables and conduct spatial regression with the mixed spatial autoregressive moving average model (*SARMA*) to assess the impacts of environmental factors on speed patterns.

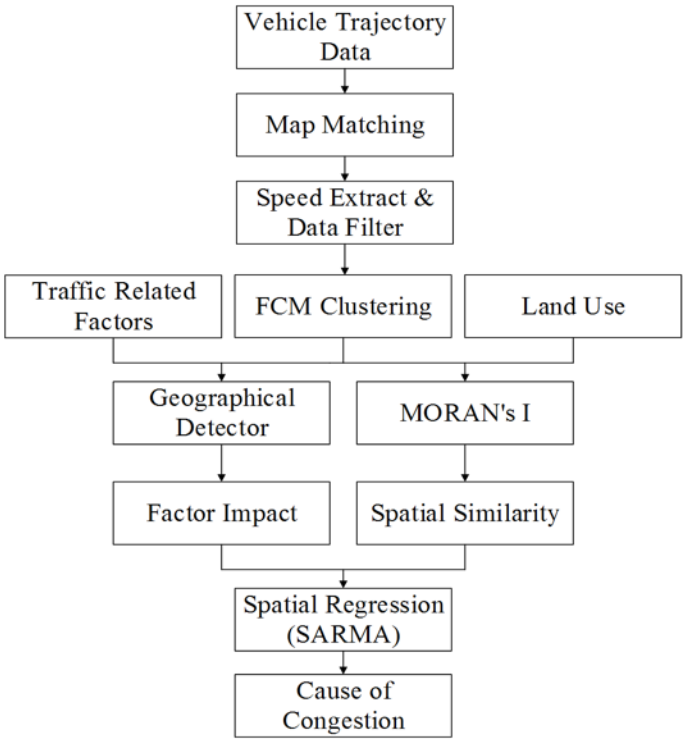



**Figure 1:** Research flow chart

The remainder of the paper is organized as follows: Section 2 introduces data collection and processing; Section 3 presents FCM clustering of road segments; and the spatial analysis models are proposed in Section 4. Finally, conclusion and future research are provided in Section 5.

## 2 Data preparation

### 2.1 Data Collection

By 2016, Shanghai has over 58,000 taxis, carrying about 4000,000 passengers each day. In this study, the Taxi FCD data on April 10, 2015, a sunny Friday with rather heavy traffic was provided by Qiangshen Company on an online open data competition (http://sodata.io/). The original dataset has 114,633,142 records, with the time interval as 30s. The original records include *Taxi ID, Status, Signal Receive Time, Signal Measured Time, Longitude, latitude, speed, etc.* (Sun, Zhang, Zhang, Chen, & Peng, 2014). Only the occupied taxi trips were kept for speed pattern generation (Cui et al., 2016) because the empty taxi could not reflect real traffic condition resulting from slowdown for searching passenger, work shifting or filling up gas during empty trips.

Focusing on the heavy traffic, only the primary and secondary roads within the Central City (Outer Ring) of Shanghai were considered. Road segment was defined as the link between two main intersections. The segments with length less than 300m were removed to avoid the excessive influence of intersections. A total of 853 road segments with traffic and built environmental data were screened out, representing the most severe congestion during the workday.

### 2.2 Data Processing

Data not matching to the nearest roads within 15m, staying still for over 5 minutes, or with speed over 120 km/h and sudden distance deviation over 1 km/min were eliminated. Thirty visits per hour were set as the threshold for segments to avoid error propagation, and 551 segments were kept for further analysis. We calculated the average speed of each segment by averaging the taxis passed through each hour. Figure 2 presents the boxplot and mean trend of the average speed on all segments by hour. The speed pattern shows obvious valleys during peak hours (7 am – 10 am & 4 pm – 5 pm), which means the most severe congestion and time-of-day congestion degree variations.

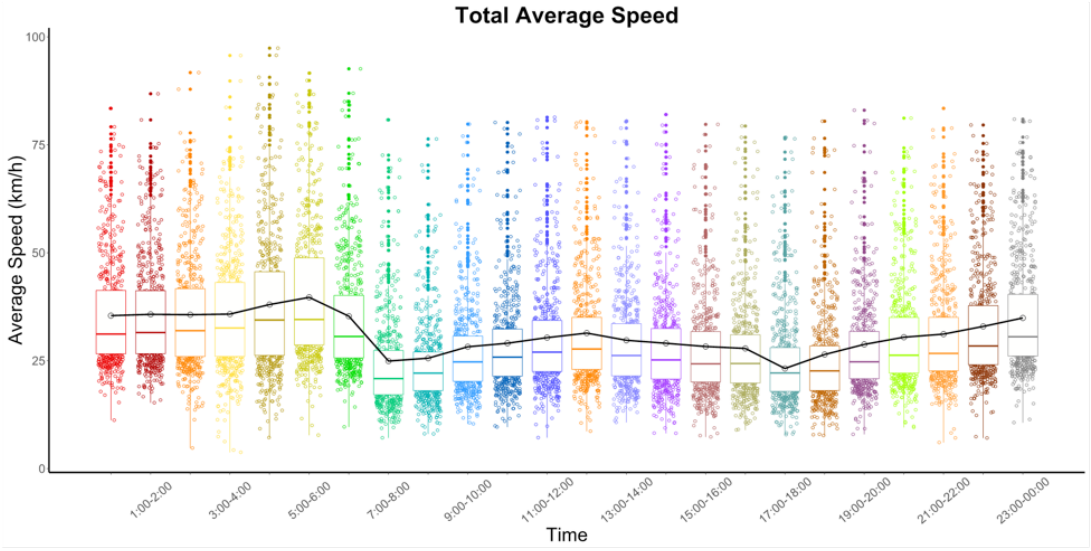



**Figure 2:** Total average speed

## 3 Clustering of road segments

### 3.1 Fuzzy C-means clustering

The 24-hour speed pattern of 551 selected segments wasas calculated, which also could be expressed by congestion. Clustering analysis, an unsupervised machine learning method, was used to aggregate road segments into groups based on their speed patterns. First, the 24-hour speed pattern of each segment was expressed with a 24-dimension vector: $y_i=\{S_1, S_2, \dots, S_{24}\}$, where $S_i$ is the average speed of time interval hour $i$.

Hard clustering (eg., K-means) has less flexibility (Azar, EI-Said, & and Hassanien, 2013; Sun &and Elefteriadou, 2012; Sun &and Elefteriadou, 2011), while the soft clustering, fuzzy C-means (FCM) clustering algorithm (Dunn, 1973; Bezdek, Ehrlich, & and Full, 1984), expresses that data points possibly belong to multiple clusters at the same time by membership degrees, which offers a much finer degree of the data model, so that the numeric results could be used for further regression. Suppose there are $N$ objects of $C$ classes ($C$ should be pre-determined). The algorithm aims to minimize the following function:

$$J(U,C) = \sum_{i=1}^{N}\sum_{j=1}^{C}(u_{ij})^m \|X_i - V_j\|_A^2$$
$$s.t. \sum_{j=1}^{C} u_{ij} = 1; 0 \le u_{ij} \le 1 \quad (1)$$

Where $X_i$ is the $i^{th}$ object, $V_j$ is the center of cluster $j$. $m$ is the fuzzifier greater than 1, and higher value means a higher degree of ambiguity. When $m$ is close to 1, it's more like hard clustering. For the best physical significance, according to Bezdek (1980), $m$ = 2 was adopted in this study. $u_{ij}$ is the grade of membership value of the $i^{th}$ object to the $j^{th}$ cluster, $\sum_{j=1}^{C} u_{ij} = 1 . \|X_i - V_j\|_A$ is the A-norm on $R^n$, measuring the similarity of objects to the assigned cluster, according to Bezdek et al. (1984). Equations (2) and (3) illustrate the iteration process to calculate the centroids of clusters and the membership values:

$$V_j = \sum_{i=1}^{N} (u_{ij})^m X_i \Big/ \sum_{i=1}^{N} (u_{ij})^m \quad (2)$$

$$u_{ij} = 1 \Big/ \sum_{s=1}^{C} \left(\frac{|X_i - V_j|}{|X_i - V_s|}\right)^{\frac{1}{m-1}} \quad (3)$$

Iteration stops when $| U_{p+1} - U_p | < \varepsilon$, where $U_p$ is the membership $u_{ij}$ at the $p^{th}$ iteration. Since the membership is within the range of 0 to 1, typically the iteration accuracy $\varepsilon$ is set as 0.001 (Bezdek et al.,, Ehrlich and Full 1984), or the iteration number is fixed at 100 (Schw &and Jensen, 2010). $\varepsilon$ was chosen as 1e-5 in this research, which is enough to both guarantee the iteration number and accuracy.

As mentioned, the cluster number is predetermined and requires revising by validity indexes to reach a meaningful and explicable result. In principle, an ideal cluster number $C$ could keep balance between inter-distance for each pair of centroids, $\|V_i - V_j\| (i \neq j)$, and intra-distance of clusters, $\|X_i - V_j\| (X_i \subseteq C_j)$. Four validity indexes appropriate for FCM were used to confirm optimal $C$:

1) **Partition coefficient:** *PC* (James, 1973) emphasizes the intensity of membership, and expressed by square weighting:

$$PC(C) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} {u_{ij}}^2 \quad (4)$$

A larger value of *PC* generally indicates a better expression of belongingness.

2) **Fuzzified PBM:** *PBMF* (Pakhira, Bandyopadhyay, & Maulik, 2005) indicates the compactness within the same cluster and the segregation between clusters. Such effect is reflected by ratio, the greater the better:

$$PBMF(C) = \frac{\max_{j,k}\{\|V_j - V_k\|\} \times E_1}{\sum_{i=1}^{N} \sum_{j=1}^{C} {u_{ij}}^m \|X_i - V_j\|} \quad (5)$$

Where, $E_1$ is a constant for a fixed sample.

3) **Minimum centroids' distance:** *MCD* (Zhu & Nandi, 2014) is the minimum distance between current cluster centers, aiming to explain the dispersion degree of clusters:

$$MCD(C) = \min_{i \neq j} | V_i - V_j |^2 \quad (6)$$

Generally speaking, *MCD* is monotone decreasing with the increase of $C$, and the suggested $C$ should be the point when the recession curve comes to stable.

4) **Fukuyama-Sugeno index:** *FSI* (Fukuyama & Sugeno, 1989) tests both the separation of all objects and the separation of clusters, and the target is to ensure these two separation degrees conforming to each other, the smaller the better:

$$FSI = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}{}^{m}(|\ X_i - V_j\ |^2 - |\ (\frac{1}{N}\sum_{k=1}^{N} X_k) - V_j\ |^2 \qquad (7)$$

These indexes evaluated the cluster number from different points of view and were calculated simultaneously.

### 3.2 Clustering result

First, the cluster number was determined through validity indexes. After testing cluster number from 2 to 7, *FSI, PBMF* and *PC* got best cluster numbers of 5, 3, 5, while for *MCD*, the trend turns to be stable when the cluster number attained 4 with an abrupt decline in centroid distance. Finally, 4 clusters would get the appropriate physical significance. Cluster 1 has 235 objects, Cluster 2 has 177 objects, Cluster 3 has 107 objects and Cluster 4 has 32 objects.

Table 1 presents the 24-hour mean, standard deviation (*SD*), coefficient of variation (*CV*) and range for the four clusters.

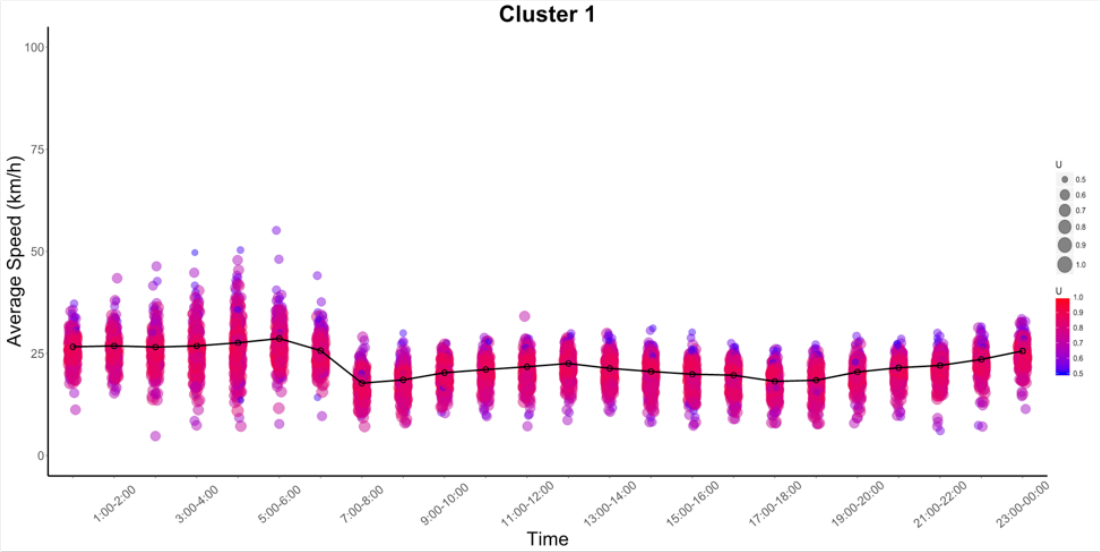
**Table 1:** Statistical indexes of each cluster

| Index | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| MEAN (km/h) | 22.59 | 29.51 | 41.11 | 61.63 |
| SD (km/h) | 3.38 | 4.23 | 6.82 | 6.27 |
| CV | 6.68 | 6.98 | 6.03 | 9.82 |
| RANGE (km/h) | 10.93 | 14.80 | 24.65 | 27.42 |

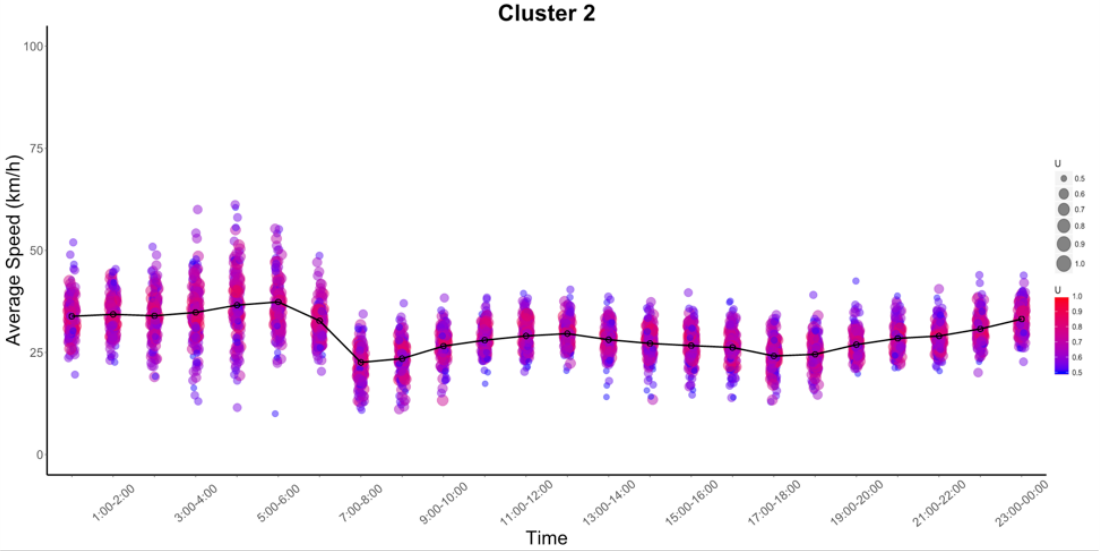
It can be figured out that Cluster 4 has the largest mean (61.63 km/h, approximate to speed limit of primary roads at 60 km/h), *CV* and range, which implies high level of service on average but also comparable larger dispersion. The mean speed of Cluster 3 is 41.11 km/h, which is approximate to speed limit of secondary road at 40 km/h. The speed related indexes experience a progressive increase trend from Clusters 1 to 4, except for the *SD* and *CV* of Cluster 3.

Figure 3 presents the 24-hour speed pattern of four clusters resulting from the FCM clustering. The temporal trajectories of road segments are plotted against the primary cluster with the highest membership values. All the scatter points are represented by a gradual change of color and size, shown in the legend. The horizontal stochastic disturbance is added for the points for better visualization, with the black polyline marking the 24-hour trend of cluster centers. As presented in Figure 3(a), Cluster 1 with the largest sample size (i.e., 235) is labeled as "Congested Segments" with the mean speed of 22.58 km/h. Its speed trajectory keeps at a low level with relative stable trend compared with other clusters. The low dispersion in Cluster 1 implies that the highly congested pattern might be caused by continuous traffic pressure, design flaw, or some certain intrinsic attributes of facilities. Cluster 2 (Figure 3(b)) is characterized by comparatively medium speed with 29.51 km/h mean, and could be labeled as "Normal Speed Segments" because its speed pattern conforms to the previous researches about typical urban road segment travel speed (Kumar & Vanajakshi, 2013). Cluster 2 has the second largest sample size, 177. Cluster 3 (Figure 3(c)) could be regarded as a critical state, and exhibits a higher speed and fluctuation changes during peak hours, thus labeled as "Unimpeded Segments". The mean speed, 41.11 km/h, is just about 7 km/h higher than that of Cluster 2, which is a comparatively small value, with the 40 km/h secondary speed limit. The other dominant factor distinguishing Cluster 3 from Cluster 2 is the peak characteristics, and these segments may be on the main commuting corridors and have a tidal phenomenon. Cluster 4 (Figure 3(d)) is labeled as "High Speed Segments", who's mean speed is the highest, 61.63 km/h, with the smallest sample size, 32. Such a continuous non-congestion pattern is rare, closed

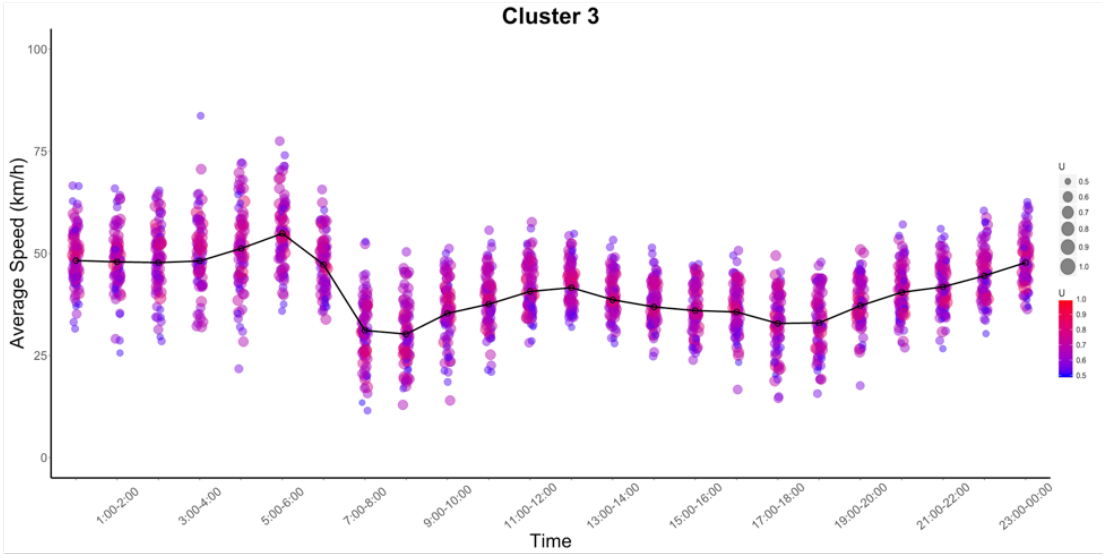
to the 60 km/h speed limit of the arterial roads in Shanghai. As the segments in Cluster 4 generally are with the highest level of service with almost no congestion, the time-of-day rationality has been proved.



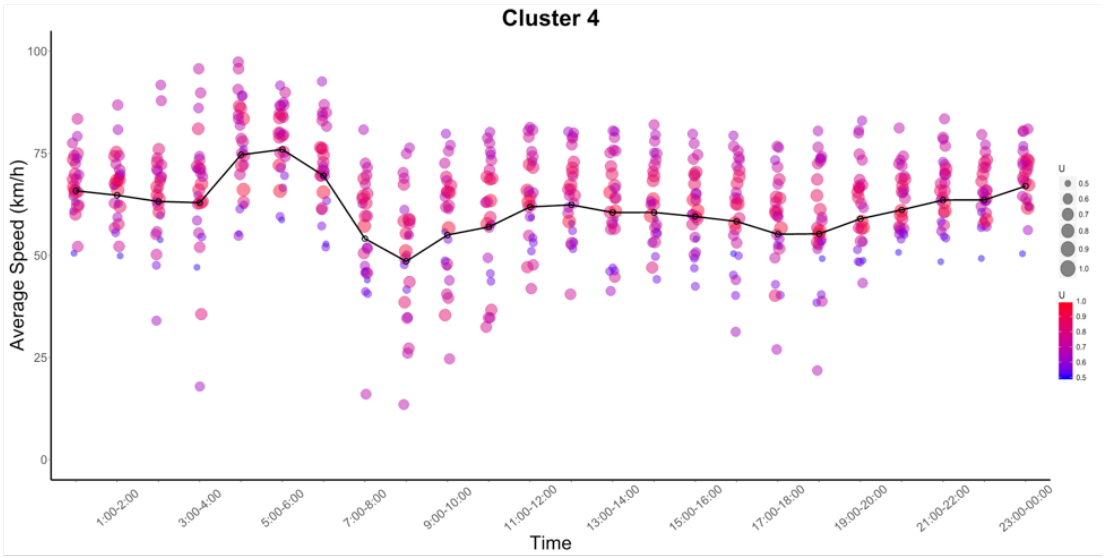

(a) Highly Congested Segments




(b) Normal Speed Segments




(c) Unimpeded Segments




(d) High Speed Segments

**Figure 3:** Temporal clustering of 24-hour speed pattern by segments

Figure 4 illustrates the spatial distribution of studied road segments based on their highest membership value. The red line is the Huangpu River, which divides Shanghai into Pudong (right) and Puxi (left). The drop marks the city center. It can be figured out that the segments in Cluster 1 concentrate mainly in Puxi, and these roads connecting the city center with outskirts, carrying a majority of commuting vehicles. This may explain why road segments in Cluster 1 have the worst traffic condition. Segments in Cluster 2 (green line) evenly distribute across the study area. Clusters 3 and 4 distribute in the surrounding parts of the study area, in other words, the roads further away from the city center have a higher probability of high speed patterns and the probability of congestion is much lower.


**Figure 4:** Clustering result by primary membership

## 4 Spatial analysis of road segments

In this research, a quantitative analysis, *SARMA*, is introduced to provide further explanation for clustering characteristics and the built environment impacts, while geographic detector and *MORAN's I* were tested ahead to applicability of *SARMA*. The explanatory variables include two categories, traffic-related factors and land use, which are widely used to explain traffic phenomenon but not in combination with taxi GPS data. For example, Zhang, Hong, Nasri, and Shen (2012) measured the impact of residential density, employment density, land-use mix, block size and distance from CBD on vehicle-miles travel. Tian et al. (2015) assessed relationship between traffic generation and mixed-use development. Briefly, main built environmental factors affecting traffic state or travel behavior could be divided into traffic-related (Hahn et al., 2002; Feng, Li, Zhao, & Hu, 2011; Zhang & Levinson, 2017) and land-use related (Wheaton, 1998; Handy, Cao, & Mokhtarian, 2005) ones.

Based on the previous researches, variables chosen for further analysis are purposed as follows:

1) Traffic-related factors:

**F1:** Road type, primary or secondary road. In the numeric analysis, 1 for the primary road and 2 for the secondary road. The average speed of the primary road was always higher than that of the secondary road in this study.

**F2:** Road segment length. Since the segment was defined as the road link between two intersections, the segment length might affect average speed greatly.

**F3:** Distance to the nearest ramp. Ramps are bottlenecks for traffic breakdown, and breakdown may affect miles away (Kerner & Klenov, 2006), so distance to the nearest ramp would impact road congestion.

**F4:** Number of bus stations along the road segment per 100m. More bus stations might reflect the commuting pressure, bring more frequent lane-changing or heavy vehicle rate (Hahn et al. 2002).

**F5:** Distance to the nearest metro station. Metro stations carry large amount of passenger flow and transship of bus or sedan.

**F6:** The relative location to the urban expressway rings. The urban expressway system in Shanghai has three rings: inner ring, middle ring and outer ring, which typically distinguish the center with the suburb. The regions divided by the three rings are donated by 1, 2, 3, 4 from the inner center to the outside of the outer ring.

**F7:** Number of parking lots open to the society within 500m per 100m. Parking lots would be the main destination for private vehicles, which may bring congestion around. Since people would walk to their final destination, parking lots' influencing scope should be the typical walking distance, 0.25 miles (Yang & Diez-Roux, 2012), and the buffer was chosen as 500m in this paper.

2) Land use:

**F8:** Number of schools within 500m per 100m. Because schools in China produce a significant traffic pressure on its surrounding roads on the time of going or leaving school by picking-up behavior of parents or school bus (Yu & Liu, 2011), it should be considered. And the influencing buffer is also 500m as parking lots.

**F9:** Distance to the nearest hospital. Hospital is the main service institution in urban areas, which is the potential congestion zones because of high traffic demand (Wen, Chin, & Lai, 2017).

**F10-12:** Land use. Land use may have significant interactions with transport (Moeckel, 2016; Miller & Evans, 2011). Land use mainly considers commercial area (**F10**), residential area (**F11**) and transportation area (**F12**), while actors like education and hospital have been involved before. Transportation land cover mainly includes the railway station, airport, transportation hubs, etc. Three factors are expressed by the area proportion within a radius of 500m:

$$P_i = {S_i}\Big/{\sum_{j=1}^{n} S_j} \tag{8}$$

Where $S_i$ is the target function area, $n$ is the number of lane use types, including commercial land, residential land, and transportation land.

Table 2 provides the minimum, mean, and maximum values of the built environmental variables. The second column gives the abbreviation of variables, and the last three columns provide a brief description of statistical data, according to the definition mentioned above.

**Table 2:** Variables in the segment cluster membership model

| Variable | Abbreviation | Min | Mean | Max |
|---|---|---|---|---|
| **F1:** Road type | Rd_type | 1 | 1.57 | 2 |
| **F2:** Road segment length (m) | Rd_len | 300 | 1314 | 4510 |
| **F3:** Distance to the nearest ramp (m) | Dist_ramp | 8.33 | 990.9 | 4304 |
| **F4:** Number of bus stations along the road segment per 100 m (stations/100m) | Num_bus | 0.0 | 0.24 | 2.47 |
| **F5:** Distance to the nearest metro station (m) | Dist_metro | 8.05 | 814.9 | 3213 |
| **F6:** Relative location to the freeway rings | Ring | 0.0 | 2 | 3 |
| **F7:** Number of parking lot with 500m | Parking | 0.0 | 3.15 | 39 |
| F8: Number of schools within 500 m per 100 m (schools/100m, r=500m) | Num_scho | 0.0 | 0.46 | 2.92 |
| **F9:** Distance to the nearest hospital (m) | Dist_hosp | 13.47 | 979.8 | 6207 |
| **F10:** Commercial area proportion (%) | Com_pro | 0.0 | 5.25 | 59.12 |
| **F11:** Residential area proportion (%) | Res_pro | 0.0 | 29.29 | 99.41 |
| **F12:** Transportation area proportion (%) | Trans_pro | 0.0 | 15.63 | 100 |

### 4.1 Geographical detector

FCM cluster classified the road segments by 24-hour speed pattern, however whether the traffic- related factors and land use have noticeable impact on congestion phenomenon remains doubtable. *Geographical detector* (Wang et al., 2010) was introduced to judge the built environmental parameters which may

be responsible for the road segments clustering. The advantage of using such geographical spatial detectors is that it considers built environmental parameters of various units. The power of determinant (*PD*) was introduced to determine whether a spatial factor may be responsible for clustering result.

Assuming there are $n$ objects, and cluster $D_i$ contains $n_{Di}$ objects, $n = \sum_{Di=1}^{4} n_{Di}$. The power of determinant factor $R$ is calculated by:

$$PD_R = 1 - \frac{1}{n\sigma_R^2} \sum_{i=1}^{4} n_{Di} \sigma_{Di,R}^2 \quad (9)$$

where $PD_R$ is the factor $R$'s power of determinant on clustering result, $\sigma_R^2$ is the global variance of factor $R$ in the study region, and $\sigma_{Di,R}^2$ is the variation of factor $R$ in cluster $Di$. Equation (9) interprets the ratio of the $n_{Di}$ weighted variation in single clusters over the global variance. The value range of $PD_R$ is [0, 1], a larger value indicates the factor $R$'s value between clusters is largely distinct, and the determinant power of $R$ is stronger. If $PD_R$ equals to 1, factor $R$ alone could perfectly classify objects.

Figure 5 presents factors' explanatory power. Bus station factor (0.130) has the highest *PD*, which means more bus stations along the road segment per 100m is related to high possibility of congestion, because bus stations of higher density reflect larger commuting volume along the road segments. The secondary factor is road type (0.105), the average speed on the primary road is always higher than that of the secondary road resulting from speed limit, intersection density, and control strategy (arterial priority). The third highest factor is the distance to the nearest hospital (0.091), which carries a high volume of patients. When the trip is concerned with diseases, people are inclined to take a taxi or private vehicle, causing higher traffic volume. The number of schools within 500m per 100m (0.084) and transportation land use (0.071) are also main explanation factors, as schools obviously attract more commuting traffic, and transportation hubs would gather huge amount of mixed traffic flow to pick up passenger or cargo. Unusual phenomenon comes from the distance to the nearest metro distance with low PD, as people using metro line mainly take public transit with less private vehicles. The factors of residential land around, segment length and location relative to freeway rings are less powerful, meaning less obvious difference between clusters, but they may be significant in a certain cluster. The low power of residential factor is particularly interesting, as our clustering result mainly reflects the overall speed pattern on a 24-hour basis while the neighborhoods may only have significant impacts on traffic in peak hours.

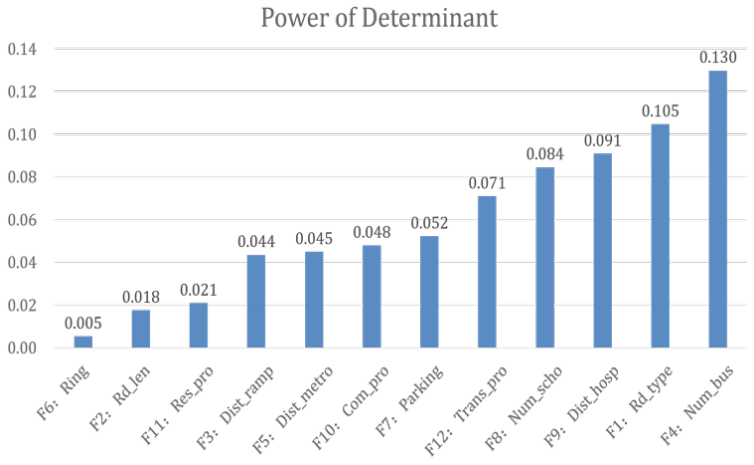



**Figure 5:** Power of each determinant in ascending sequence

By the *geographical detector,* some factors were found affecting congestion formation in all clusters,

which disclose global influence. Others may only have exclusive impacts on certain clusters with unobvious global influence (low *PD*), which may require further investigation for each cluster, respectively.

### 4.2 MORAN's I

*GLOBAL MORAN's I* (Moran, 1950) measures spatial correlation and tests whether observed objects have similarities with the spatial adjacency objects. For the value of *MORAN's I* ranges, [-1, 1], $I = 0$ means totally spatial independence, $I > 0$ reflects positive correlation and $I < 0$ means negative correlation. The calculation formula is as follows:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(X_i-\bar{X})(X_j-\bar{X})}{(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})\sum_{i=1}^{n}(X_i-\bar{X})^2} \quad (i \neq j) \tag{10}$$

where, $X_i$ and $X_j$ are the observed values, which indicate membership to clusters, and $\bar{X}$ is the mean value. $W_{ij}$ is the spatial weight matrix describing the spatial relationship among objects. *MORAN's I* can be calculated cluster by cluster.

Spatial weight matrix plays an important role in the spatial analysis. Binary joint matrix is commonly used to characterize spatial weight matrix (Cliff &and Ord, 19821), if two observations directly connect with each other, $W_{ij} = 1$, otherwise $W_{ij} = 0$. However, binary joint matrix is not suitable for line object, such as road segments. Because of the connectivity of roads and the transmissibility property, road segments would impact each other. The spatial weight matrix used was based on distance decay (Greicius, Krasnow, Reiss, & Menon, et al. 2003), and midpoint of the road segment represents the location feature:

$$w_{ij} = \begin{cases} \exp\left[-0.5\left({}^{d_{ij}}/_{b}\right)^2\right], & d_{ij} < b \\ 0, & otherwise \end{cases} \tag{11}$$

Where $b$= 1000m, and the matrix is standardized by row, $d_{ij}$ is the distance between midpoints of two road segments.

*Z test* was used to access the result of *MORAN's I*, and it could be interpreted by typically *P-value* (Cliff & Ord, 1982):

$$Z = {}^{[1-E(I)]}\Big/_{\sqrt{VAR(I)}} \tag{12}$$

Where *E(I)* is the mean value of *MORAN's I*, and *VAR*(*I*) is the variation of *MORAN's I*.

*MORAN's I* evaluated whether objects' memberships to a cluster have aggregation effects spatially, the result is shown in Table 3.

**Table 3:** *MORAN's I* for clusters

| Index | Cluster1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| *MORAN's I* | 0.246 | 0.032 | 0.14 | 0.271 |
| *Z value* | 5.997 | 0.816 | 3.434 | 6.694 |
| *P value* | 1.01e-09 | 0.207 | 0.0 | 1.09e-11 |

All *MORAN's I* values are larger than 0, and only Cluster 2 fails to pass the 5% level of the significance test. The positive value indicates the bipolar aggregation phenomenon. As in a certain cluster, road segments with higher membership gather together and lower ones as well, indicating lag effect for neighboring segments, which means neighboring segments have similar possibility and level of congestion. Although *MORAN's I* value is still comparable small, it discloses spatial gathering and explains the

effectiveness of FCM clustering.

### 4.3 Spatial regression of road segments

The results from FCM are used to conduct multiple regression analysis based on continuous membership $u_{ij}$ and environment factors. *Geographical detector* and *MORAN's I* have proved the factors' impact on clustering result and the spatial similarity of the nearby road segments, so in addition to the 12 factors, lagging influence from neighboring segments are also considered. A spatial model involving both spatial autocorrelation and multivariable system is therefore preferred. A spatial lag mode called mixed spatial autoregressive moving average model (*SARMA*) (Anselin, Bera, Florax, & Yoon, et al. 1996) is introduced to consider both dependence and errors with nearer objects having a greater impact. The structure of *SARMA* is shown in Equations (13) & (14).

$$y_i = (1 - \rho W)^{-1} X\beta + (1 - \rho W)^{-1} u \quad (13)$$

$$u = (1 - \lambda W)^{-1} \varepsilon, \varepsilon \sim N(0, \sigma^2 I) \quad (14)$$

where $y_i$ is the segment's membership belonging to Cluster $i$, $X$ is the vector of environment characteristics; $\rho$ is the spatial autoregressive parameter measuring neighborhood effects, $\rho$>0 means positive correlation and vice versa; $\lambda$ is the spatial error coefficient, disclosing and quantifying the inherent similarity or dissimilarity; $\varepsilon$ is the random error term; $W$ is the spatial weight matrix mentioned in Section 4.2, and $\beta$ is the coefficient vector.

Before *SARMA* regression, each factor was standardized by Equation (15) to make the estimated coefficient at a comparable magnitude:

$$y_i' = {[y_i - E(y)]}/{SD(y)} \quad (15)$$

Where, $y_i'$ is the standardized value, E(y) is mean and SD(y) is the standard deviation.

Results of the *SARMA* regression is presented in Table 4, in which strong factor influence and spatial lagging effect have been disclosed. In other words, surrounding location characteristics and neighboring road segments are related with the type of speed pattern on the road segment.

For **Cluster 1 (Highly Congested Segments),** all significant factors show a continuous traffic pressure. The merging area on a ramp may cause congestion, thus blocking the ground road. Bus stations and parking lots alongside also bring continuous traffic flow. For road segments whose highest membership belonging to Cluster 1, 55.1% of the schools around are universities or vocational schools, seldom reflecting commuting feature compared with high schools. Moreover, demand for hospitals is general stupendously high, stimulating private traffic flow and taxis. Noting the coefficient for road type is positive, which indicates that secondary road with lower speed limit causes more impact. An interesting finding is that higher portion of transportation type land -use lowers the membership degree for Cluster 1, which means less congestion. This is probably due to the fact that the high proportion of transportation-type land use represents transit hub, such as an airport or a railway station, where road network in the vicinity is usually well organized. Small hubs such as highway bus stations or logistic stations are generally accompanied with congested high occupancy traffic.

**Cluster 2 (Normal Speed Segments)** has a strong trend of commuting traffic demand. Significant factors with commuting phenomenon including schools, metro stations and commercial land-use effects. However, the influential factors such as bus station, parking lot and hospital in the vicinity come to be less significant, and the average speed is improved overall. Longer road segment generally means

fewer signal controls, which is also a critical factor. And ramps nearby will greatly affect the travel speed, which causes more congestion.

**Cluster 3 (Unimpeded Segments)** has less significant factors, and the factors are inversed compared with Cluster 1. Road segments in Cluster 3 are mainly primary roads (with negative coefficient), and the disturbing factors, such as ramps, bus stations, and parking lots become fewer or further, while the proportion of transportation land-use turns to be extremely high. These actually relieve the traffic pressure on the road.

For **Cluster 4 (High Speed Segments)**, primary-membership objects' mean speed is close to 60km/h speed limit, while the surrounding built environmental factors are similar to the ones in suburban, which can also be obtained from Figure 4. The low density of ramp, bus station, metro station and hospital all prove this conclusion. According to this result, relative location to the expressway rings and the proportion of residential have little impact on the congestion formation. Based on general understanding, residential area is always regarded as the origin of commuting traffic flow and causes congestion. This may be true in peak hours. However, when considering the entire 24-hour patterns, the impact of residential district may not be obvious. Furthermore, commercial land, such as CBD only plays a key role in Cluster 2, and contributes little to road clustering analysis. However, it acts as an important threshold between "Highly Congested" and "Normal Speed."

The spatial-lagged dependent variable $\rho$ and spatial error variable $\lambda$ were chosen and further analyzed. Spatial-lagged dependent variable indicates the contagious or alien of a dependent variable based on positive or negative values. For Clusters 1, 2 and 3, $\rho$ is significantly positive, indicating that the adjacent road segments have a similar cluster type. This discloses the fact that for segments within 1000 m, their congestion pattern and traffic condition are mainly driven by the environment factors around. However, $\rho$ value for Cluster 4 is negative, indicating low speed on road segments near a high speed segment. This may result from dis-connectivity of roads or high traffic demand in a special location. The significance of $\lambda$ associated with Clusters 2 and 3 is negative, which implies that unobserved factors nearby impact membership differently. While $\lambda$ for Cluster 4 is positive, showing unobserved neighboring variables have parallel effects on segment clustering.

**Table 4:** SARMA models for cluster membership prediction

| Factors | Cluster 1<br>Highly Congested | Cluster 2<br>Normal Speed | Cluster 3<br>Unimpeded | Cluster 4<br>High Speed |
|---|---|---|---|---|
| (Intercept) | 0.3731 *** | 0.2885 *** | 0.1332 *** | 0.0803 *** |
| | (0.0328) | (0.0245) | (0.0138) | (0.0105) |
| Rd_type | 0.0695 *** | | -0.0577 *** | -0.0181 *** |
| | (0.0113) | | (0.0083) | (0.0059) |
| Rd_len | | 0.0229 ** | | |
| | | (0.0098) | | |
| Dist_ramp | -0.0192 * | -0.0306 *** | 0.0165 ** | 0.0395 *** |
| | (0.0114) | (0.0085) | 0.0073 | (0.0084) |
| Num_bus | 0.0831 *** | 0.0186 * | -0.0369 *** | -0.0173 *** |
| | (0.0114) | (0.0096) | (0.0085) | (0.0057) |
| Dist_metro | | -0.0229 ** | | 0.0257 *** |
| | | (0.0099) | | (0.0080) |
| Ring | | | | |
| Parking | 0.0305 ** | | -0.0160 * | |
| | (0.0140) | | (0.0089) | |
| Num_scho | 0.0373 *** | -0.0269 ** | | |
| | (0.0139) | (0.0106) | | |
| Dist_hosp | -0.0307 ** | | | 0.0227 *** |
| | (0.0140) | | | 0.0082 |
| Com_pro | | 2.3256 ** | | |
| | | (1.1483) | | |
| Res_pro | | | | |
| Trans_pro | -3.6390 *** | | 0.0239 *** | |
| | (1.2078) | | (0.0085) | |
| Rho | 0.1242 ** | 0.1432 * | 0.2978 * | -0.2764 * |
| | (0.0789) | (0.0767) | (0.0716) | (0.0882) |
| Lambda | | -0.1979 ** | -0.2465 * | 0.4858 *** |
| | | (0.0915) | (0.0894) | (0.0684) |
| LR test | 9.9031 *** | 3.3252 * | 9.5079 *** | 31.497 *** |
| Log likelihood | - | 55.7193 | 113.7466 | 296.686 |
| | -37.7427 | | | 8 |
| AIC | 107.49 | -79.44 | -195.49 | -561.37 |

Note: * indicates p<0.1, ** indicates p<0.05, *** indicates p<0.01. Standard errors are recorded in parentheses.

The regression result could be implemented to recognize the congested road segments in urban area of Shanghai. During the process of transportation planning or urban planning, the result could also be applied to assess road network layout and its combination with land use and traffic-related factors. For example, a secondary road segment with low proportion of transportation and high density of bus station has higher probability of suffering continuous congestion, such as Cluster 1. With better insights for problematic road segments, some spatial or temporal redesigns, such as setting variable lanes, widening roads, optimizing road function or setting bus transit lane could be conducted.

## 5 Conclusion and recommendations

In this study, taxi GPS data were used to analyze 24-hour speed pattern of primary roads and secondary roads. Speed was used as the main indicator to further disclose the congestion phenomenon of roads. Correlations with 12 built environmental factors including traffic-related factors and land use were further investigated using the data from Shanghai, China as a case study.

First, the average speed of road segments per hour was extracted from GPS trajectories. Fuzzy C-means algorithm was applied to cluster the segments with 24-hour dimension vector based on average speed to classify roads into 4 different congestion level. A *geographical detector* was then utilized to find key common factors related to congestion patterns. *MORAN's I* was computed based on types of cluster to investigate spatial similarity of adjacent segments and confirmed a spatial lagging effect. Based on previous findings, a spatial regression model was implemented to identify influential environmental factors associated with each cluster, and the interaction between neighboring segments.

Compared with previous studies, this paper combined clustering method with quantitative spatial analysis for better explanation. The influencing factors of congestion were explored using spatial regressions, thus to provide better understanding of existing congestion level and quick service evaluation based on environmental data and road conditions.

While the results are promising, further studies need to be conducted to improve the performance of the model. First, in this research, the taxi GPS data doesn't cover the entire scope of Shanghai, confining the study mainly to urban areas. Secondly, only one weekday data was analyzed, while multi-date analysis may have to be carried out in the future. Particularly during the spatial analysis, the proposed models failed to consider the interactions of environmental factors, which may ignore significant impacts. Moreover, certain research for peak hour analysis would also be carried out in the future, since short-period analysis may disclose some important features or distinct phenomenon only appearing in peak duration. Since peak hours generally have more variation and may need additional accuracy, it would also be interesting to have various temporal divisions for data, which may result in more realistic classifications with respect to congestions. An attempt would be that based on general statistics of the current division, further dividing the congested hours into half-hour period considering the ordinary taxi trip would be less than 30 minutes (or even 15 minutes), while using 1 hour for non-peak trips.

## Acknowledgements

This work was supported in part by the Humanities and Social Science Research Project, Ministry of Education, China [15YJCZH148], the Philosophy and Social Science Research Project of Shanghai, China [2014BGL009], and the Fundamental Research Funds for the Central Universities in China [15JCZZ04]. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsors.

## References

Anselin, L., Bera, A. K., Florax, R., & Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics, 26*(1), 77–104.

Azar, A. T., El-Said, S. A., & Hassanien, A. E. (2013). Fuzzy and hard clustering analysis for thyroid disease. *Computer Methods and Programs in Biomedicine, 111*(1), 1–16.

Azimi, M., & Zhang, Y. (2010). Categorizing freeway flow conditions by using clustering methods. *Transportation Research Record, 2173,* 105–114.

Bezdek, J. C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Tansactions on Pattern Analysis and Machine Intelligence, 2*(1), 1–8.

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences, 10*(2–3), 191–203.

Chen, C., Zhang, D., Li, N., & Zhou, Z. H. (2014). B-Planner: Planning bidirectional night bus routes using large-scale taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems, 15*(4), 1451–1465.

Cliff, A. D., & Ord, J. K. (1982). Spatial processes: Models and applications. *Quarterly Review of Biology, 57*(2).

Cui, J., Liu, F., Janssens, D., An, S., Wets, G., & Cools, M. (2016). Detecting urban road network accessibility problems using taxi GPS data. *Journal of Transport Geography*, *51*, 147–157.

Ding, J., Gao, S., Jenelius, E., Rahmani, M., Huang, H., Ma, L., & Ben-Akiva, M. (2014). Routing policy choice set generation in stochastic time-dependent networks: Case studies for Stockholm, Sweden, and Singapore. *Transportation Research Record, 2466,* 76–86.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics, 3*(3), 32–57.

Feng, H., Li, C., Zhao, N., & Hu, H. (2011). Modeling the impacts of related factors on traffic operation. *Procedia Engineering, 12,* 99–104.

Fukuyama, Y., & Sugeno, M. (1989). A new method of choosing the number of clusters for fuzzy C-means method. Presented at the 5th Fuzzy System Symposium, Kobe, Japan.

Goddard, J. B. (1970). Functional regions within the city center: A study by factor analysis of taxi flows in central London. *Transactions of the Institute of British Geographers, 49,* 161–182.

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences, 100*(1), 253–258.

Hahn, E., Chatterjee, A., Younger, M. S., Hahn, E., Chatterjee, A., & Younger, M. S. (2002). Macro-level analysis of factors related to area-wide highway traffic congestion. *Transportation Research Record,* 1817, 11–16.

Handy, S., Cao, X., & Mokhtarian, P. (2005). Correlation or causality between the built environment and travel behavior? Evidence from northern California. *Transportation Research Part D Transport and Environment, 10*(6), 427–444.

He, F., Yan, X., Liu, Y., & Ma, L. (2016). A traffic congestion assessment method for urban road networks based on speed performance index. *Procedia Engineering*, 137, 425¬–433.

Hu, X., An, S., & Wang, J. (2014). Exploring urban taxi drivers' activity distribution based on GPS data. *Mathematical Problems in Engineering, 2014*(2), 1–13.

Hwang, K., Wu, K., & Jian, R. J. (2006). Modeling consumer preference for Global Positioning System-based taxi dispatching service: Case study of Taichung City, Taiwan. *Transportation Research Record,1971,* 99–106.

Jiménez-Meza, A., Arámburo-Lizárraga, J., & Fuente, E. D. L. (2013). Framework for estimating travel time, distance, speed, and street segment level of service (los), based on GPS data. *Procedia Technology, 7*(4), 61–70.

Kerner, B. S., & Klenov, S. L. (2006). Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: Congestion nucleation in spatially non-homogeneous traffic. *Physics, 1965*, 473–492.

Kumar, V., & Vanajakshi, L. D. (2013). Modewise travel time estimation on urban arterials using transit buses as probes. Paper presented at the 92nd Annual Meeting of the Transportation Research Board, Washington, D.C.

Lu, Y., & Li, S. (2014). An empirical study of with-in day OD prediction using taxi GPS data in Singapore. *Langmuir the Acs Journal of Surfaces and Colloids, 30*(31), 9567–9576.

Miller, J. S., & Evans, L. D. (2011). Divergence of potential state-level performance measures to assess transportation and land use coordination. *Journal of Transport and Land Use*, *4(*3), 81–103.

Moeckel R. (2016). Constraints in household relocation: Modeling land-use/transport interactions that respect time and monetary budgets. *Journal of Transport and Land Use, 10*(1), 211–228.

Montero, L., Pacheco, M., Barcelo, J., Homoceanu, S., & Casanovas, J. (2016). A case study on cooperative car data for traffic state estimation in an urban network. Presented at the 95th Annual Meeting of the Transportation Research Board, Washington, D.C.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37(*1–2), 17–23.

Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2005). A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets and Systems*, *155*(2), 191–214.

Pan, G., Qi, G., Wu, Z., Zhang, D., & Li, S. (2013). Land-use classification using taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems 14*(1), 113–123.

Qian, X., & Ukkusuri, S. V. (2015). Exploring spatial variation of urban taxi ridership using geographically weighted regression. Paper presented at the 94th Annual Meeting of the Transportation Research Board, Washington, DC.

Qing, C., Parfenov, S., & Kim, L. J. (2015). Identifying travel patterns during extreme weather using taxi GPS data. Presented at Transportation Research Board 94th Annual Meeting, Washington, DC.

Schw, M. V., & Jensen, O. N. (2010). A simple and fast method to determine the parameters for fuzzy C–means cluster analysis. *Bioinformatics*, *26*(22), 2841–2848.

Sun, D., & Elefteriadou, L. (2011). Lane changing behavior on urban streets: A focus group based study. *Applied Ergonomics: Human Factors in Technology and Society, 42*(5), 682–691.

Sun, D., & Elefteriadou, L. (2012). Lane changing behavior on urban street: An "in-vehicle" field experiment based study. *Computer-Aided Civil and Infrastructure Engineering*, *27*(7), 525–542.

Sun, D., Zhang, C., Zhang, L., Chen, F., & Peng, Z. R. (2014). Urban travel behavior analyses and route prediction based on floating car data. *Transportation Letters*, *6*(3), 118–125.

Tang, L., Yang, X., Kan, Z., & Li, Q. (2015). Lane-level road information mining from vehicle GPS trajectories based on Naïve Bayesian Classification. *ISPRS International Journal of Geo-Information, 4*(4), 2660–2680.

Tang, J., Jiang, H., Li, Z., & Li, M. (2016). A two-layer model for taxi customer searching behaviors using GPS trajectory data. *IEEE Transactions on Intelligent Transportation Systems, 17,* 1–7.

Tian, G., Ewing, R., White, A. Hamidi, S., Walters, J., & Goates, J. P. (2015). Traffic generated by mixed-use developments: Thirteen-region study using consistent measures of built environment. *Journal of Urban Planning and Development, 137*(3), 248–261.

Tulic, M., Bauer, D., & Scherrer, W. (2014). Link and route travel time prediction including the corresponding reliability in an urban network based on taxi floating car data. *Transportation Research*

*Record, 2442*, 140–149.

Wang, J. F., Li, X. H., Christakos, G., Liao, Y. L., Zhang, T., Gu, X., & Zheng, X. Y. (2010). Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *International Journal of Geographical Information Science, 24*(1), 107–127.

Wang, H., Peng, Z. R., Lu, Q. C., Sun, J., & Bai, C. (2017). Assessing effects of bus service quality on passengers' taxi-hiring behavior. *Transport.* Advance online publication. doi: 10.3846/16484142.2016.1275786

Wen, T. H., Chin, W. C., & Lai, P. C. (2017). Understanding the topological characteristics and flow complexity of urban traffic congestion. *Physica A: Statistical Mechanics and its Applications, 473*(1), 166–177.

Wheaton, W. C. (1998). Land use and density in cities with congestion. *Journal of Urban Economics, 43*(2), 258–272.

Yang, Y., & Diez-Roux, A. V. (2012). Walking distance by trip purpose and population Subgroups. *American Journal of Preventive Medicine, 43*(1), 11–19.

Yazici, M. A., Kamga, C., & Singhal, A. (2016). Modeling taxi drivers' decisions for improving airport ground access: John F. Kennedy airport case. *Transportation Research Part A: Policy and Practice, 91,* 48–60.

Yu, J., & Lu, P. (2016). Learning traffic signal phase and timing information from low-sampling rate taxi GPS trajectories. *Knowledge-Based Systems, 110,* 275–292.

Yu, L., & Liu, Y. (2011). Traffic characteristics analysis and suggestions on school bus operation for primary school students in Beijing. *Journal of Transportation Systems Engineering & Information Technology, 11*(5), 193–200.

Zhang, J., Qiu, P., Duan, Y., Du, M., & Lu, F. (2015). A space-time visualization analysis method for taxi operation in Beijing. *Journal of Visual Languages and Computing, 31,* 1–8.

Zhang, L., & Levinson, D. (2017). A model of the rise and fall of roads. *Journal of Transport and Land Use, 10*(2), 1–23.

Zhang, L., Hong, J. H., Nasri, A., & Shen, Q. (2012). How built environment affects travel behavior: A comparative analysis of the connections between land use and vehicle miles traveled in U.S. cities. *Journal of Transport and Land Use, 5*(3), 40–52.

Zhu, Z., & Nandi, A. K. (2014). Blind digital modulation classification using minimum distance centroid estimator and non-parametric likelihood function. *IEEE Transactions on Wireless Communications, 13*(8), 4483–4494.